WHAT IS CLAIMED IS:

1. A method for selecting a server from a plurality of servers to service a request for content, comprising:

designating a director from the plurality of servers to receive the request, wherein the designation is made on a request-by-request basis; and

allocating to the director the task of selecting a server to service the request from the plurality of servers, said server having stored thereon the content, the director using a state table comprising parametric information for servers in the plurality of servers, wherein said parametric information comprises information identifying assets maintained on each server in the plurality of servers.

2. The method of claim 1, wherein the step of designating comprises designating the director in a round-robin fashion.

3. The method of claim 1, wherein the step of designating comprises designating the director on the basis of lowest load.

4. The method of claim 1, wherein the step of selecting further comprises selecting the director if the content is present on the director.

5. The method of claim 1, wherein said parametric information further comprises functional state and current load of each server.

6. The method of claim 1, wherein said parametric information further comprises whether each server comprises extended memory.

7. The method of claim 1, wherein said parametric information further comprises whether each server comprises an inline adaptable cache.

8. The method of claim 1, wherein said parametric information further comprises whether each asset is a new release.

9. The method of claim 1, further comprising rejecting the request if the content is not present on any of the plurality of servers.

10. The method of claim 1, further comprising forwarding the request to the selected server.

11. The method of claim 1, further comprising redirecting the request to the selected server.

12. The method of claim 1, wherein the step of selecting further comprises:

calculating a load factor for each server in the plurality of servers having the content;

identifying as available servers one or more servers whose parameters are below threshold limits;

selecting a server from the available servers having the lowest load factor; and

otherwise selecting a server having the lowest load factor from the plurality of servers having the content.

13. A server for directing a request for content among a plurality of servers comprising:

a state table comprising parametric information for each server in the plurality of servers, said parametric information comprising information identifying assets maintained on the plurality of servers; and

a communication component for sending changes to the state table to the plurality of servers.

14. The server of claim 13, wherein the server is a member of a load-balancing group, and the communication component sends changes to servers in the load-balancing group.

15. The server of claim 13, further comprising a redirection means for acknowledging the client request and identifying one of the plurality of servers where the requested asset is stored.

16. The server of claim 13, further comprising a forwarding means for sending the client request to one of the plurality of servers where the requested asset is stored.

17. The server of claim 13, wherein said parametric information further comprises functional state and current load of each server.

18. The server of claim 13, wherein said parametric information further comprises whether each server comprises extended memory.

19. The server of claim 13, wherein said parametric information further comprises whether each server comprises an inline adaptable cache.

20. The server of claim 13, wherein said parametric information further comprises whether each asset is a new release.